图书情报工作
第62卷 第15期　2018年8月

基于引文内容分析的引用情感识别研究*

■ 廖君华[1]　刘自强[2,3]　白如江[1]　陈军营[1]
[1] 山东理工大学科技信息研究所　淄博 255049　[2] 中国科学院成都文献情报中心　成都 610041
[3] 中国科学院大学　北京 100190

摘要：［**目的/意义**］针对自动识别论文引用情感问题，提出一种基于引文内容分析的识别方法并进行可视化展示，克服基于简单引用频次计量无法区分不同引用情感的问题。［**方法/过程**］首先，利用正则表达式抽取出论文全文中的引文内容信息；然后，利用 TF-IDF 算法筛选出引用情感特征词，结合情感词典，利用情感分析技术对引文内容进行引用情感识别；最后，利用可视化工具展示出引用情感整体分布情况。［**结果/结论**］该方法能够有效识别出抗衰老领域论文数据集中引用情感情况。实验结果显示，该领域正面引用占总引用次数的21%，中立引用占总引用次数的78%，负面引用仅占总引用次数的1%。与传统引文网络相比较，基于引用情感的可视化图谱可以有效识别出不同引用情感在整体数据集合上的分布情况。

关键词：引文内容分析　引用情感　情感分析　可视化
分类号：G250
DOI：10.13266/j.issn.0252－3116.2018.15.013

1　引言

引文分析（Citation Analysis，CA）对面向科技创新的战略情报研究和服务工作起着重要的作用。其中，基于期刊论文被引次数、H指数和影响因子等指标的引文分析方法广泛应用于科学知识评价、科学发展模式揭示和科学前沿探测等。但是，由于引用行为和引用情感的复杂性，以被引用次数为基础的传统引文分析方法存在一定的不足：传统的引文分析方法将引文等同看待，并且忽略了引文在文献中的具体表现，比如引用位置、引文次数和引用语境等信息；仅仅通过被引用次数无法揭示出施引文献和被引文献在研究内容上的关联，在一定程度上降低了引用分析的准确性和有效性[1]。此外，随着科研评价体系的不断发展对引文分析提出了新的要求，比如：国家自然科学奖规定“得到国内外自然科学界公认是指其重要科学结论已为国内外同行在重要国际学术会议、公开发行的学术刊物，尤其是重要学术刊物以及学术专著所正面引用或者应用”[2]。所以，如何有效识别、判断“正面、负面引用等”引用情感倾向，改进基于引用次数的引文分析评价方法，有待进一步深入研究。

随着文本挖掘、自然语言处理和可视化等技术方法不断进步，逐渐可以利用其进行提取、计算和挖掘隐藏在引文内容中的语义信息与关联。此外，随着开放存取（Open Access，OA）运动以及结构化全文数据库的建设、发展，将引文分析研究推向一个新阶段－“引文内容分析（Citation Content Analysis，CCA）”，并逐渐获得越来越多研究者的关注与认可，应用于引用情感、引文主题识别等研究领域[3]。

2　相关研究

引文内容分析相关研究早在20世纪70年代就已出现。M. J. Moravcsik 等通过对引文内容及其上下文进行解读，细致分析了引用情感倾向、引用作用和重要程度，其研究表明了引文内容分析的必要性[4]。进而，H. Small 通过人工判读、归纳总结的方法，分析了化学学科高被引论文的引用内容，认为引文内容是施引文献中观点表达的概念符号。随着引文内容分析研究的展开，研究人员尝试结合引文内容分析改进基于引用

* 本文系教育部人文社会科学研究青年基金项目“基于引文内容分析的科技创新路径识别研究”（项目编号：16YJC870008）和山东理工大学高等教育研究项目（项目编号：2018GJY08）研究成果之一。
作者简介：廖君华（ORCID：0000－0002－8641－0080），讲师，硕士；刘自强（ORCID：0000－0003－1814－8655），硕士研究生；白如江（ORCID：0000－0003－3822－8484），副研究馆员，博士，通讯作者，E-mail：brj@sdut.edu.cn；陈军营（ORCID：0000－0003－3550－1641），硕士研究生。
收稿日期：2018－01－21　修回日期：2018－05－03　本文起止页码：112－121　本文责任编辑：杜杏叶

次数的引文分析方法[5]，1980 年，H. Small 等提出了结合引文内容分析的同被引聚类分析方法，首先基于同被引聚类分析某学科领域的演化过程，然后通过引文内容分析，利用主题词、短语概括表征引文具体内容，进而分析同被引聚类的主题，揭示共被引文献之间的深度关联，提高共被引链接的认知价值。并通过对重组 DNA 领域的实证研究，证明该方法为探索学科领域的发展演化具有重要意义[6]。但是，由于当时期刊数据库全文质量以及计算机技术的限制，研究人员主要采用人工判读、归纳总结的方法进行引文内容分析，难以处理大样本数据而且人工判读主观性较强，因此准确性受到一定的质疑，限制了引文内容分析的进一步发展。

随着自然语言处理技术和全文数据库的发展，引文内容分析获得了新的发展，并对传统引文分析的发展注入了新的活力。2012 年，Y. Ding 等提出了引文内容分析（Citation Content Analysis，CCA）研究框架，并指出引文内容分析是下一代引文分析的方向，能够拓展和深化引文分析的研究与应用[7]。祝青松、冷伏海等以碳纳米管领域的高被引论文为研究对象，进行引文内容挖掘、分析，利用 C-value 算法识别出引文内容中的研究主题，其研究表明与基于标题、摘要等字段的主题识别结果相比，基于引文内容分析的主题识别结果与论文研究内容更加符合，能较好地揭示被引文献和施引文献之间在语义内容上的关联，认为引文内容分析是对传统以被引用次数为基础的引文分析的重要补充[8]。陆伟等指出为更好地支持文献语义关系挖掘，将自然语言处理、机器学习技术引入引文内容分析，并提出了一套引文内容标注框架[9]。赵蓉英等认为引文内容分析是引文分析的新发展，可以更加准确地测度和评价被引作者、期刊影响力，透视作者的引证动机等，对科学计量学和科学学的发展大有裨益[10]。在此基础上，赵蓉英等于 2016 年结合引文内容分析方法，提出了基于位置的共被引分析框架，证明了结合引文内容分析的基于位置的共被引方式明显优于传统共被引分析方法[11]。

在引用情感类型研究方面，早在 1962 年，E. Garfield 就发现了引文频次分析的不足，指出引用情感的多样性，并归纳出向开拓者致敬、向同行致敬等 15 种引用情感[12]。此后，V. Cano 等专家学者也指出由于引用行为是复杂多样的，引用情感并不总是正面的（还存在负面引用、虚假引用等），简单的被引频次并不足以衡量学术影响力的高低[13－17]。因此，准确、高效地识别引用情感倾向，判断正面、负面引用，可以有效提升基于简单引用频次的评价质量。

在引用情感识别方法方面，M. J. Moravcsik 等通过人工判读引文全文对引用情感进行了研究，并将引用情感分为肯定引用、否定引用等 5 个维度[4]；同年，在 M. J. Moravcsik 研究的基础上，D. E. Chubin 等研究引文内容分析对于引用著录分析的辅助、替代作用，其中，利用分类树进行引用情感倾向分析，将分类树的第一层分为正面、负面引用两个子树[18]。总体来说，由于信息技术发展的限制，引用情感识别主要利用问卷调查分析和人工判读引文内容两种方法，存在效率低、主观性较强等不足。

近年来，基于自然语言处理技术的引用情感识别相关研究获得了一定发展。S. Teufel 等提出了一种基于监督式机器学习的引用情感自动分类方法框架，利用情感分析技术进行引用功能（引用情感）分类，具体分为不足、肯定、对比和中立 4 个类别，指出利用情感分析技术能够准确、有效识别引用情感[19]。刘盛博等提出了一种基于数据挖掘技术的引用情感识别方法，以 PubMed 全文数据库为数据来源，利用引用内容语义结构与特征词来判断引用情感（正面引用、负面引用和中立引用），并以此为基础，构建了一个基于引用内容的引文评价平台[20]。基于自然语言处理技术的引用情感识别，相较于利用调查问卷、人工判读分析方法，能够提高分析效率与客观性，存在的不足是引用情感识别结果分析较为浅显，如何有效利用引用情感识别结果有待进一步深入研究。

综上所述，传统的引用情感识别方法主要利用问卷调查分析和人工判读引文内容两种方法，存在效率低、主观性较强等不足；基于统计自然语言处理技术的引用情感识别方法难以有效分析引用情感识别结果，在一定程度上降低了引用情感识别的准确性和有效性。特别是在特征词库构建方面，前期研究工作主要采用已有的感情词典进行情感极性判别，而科技论文引用情感与通用情感词典会有较大差异。因此，为了改进目前引用情感识别相关研究中缺乏有效的情感词典构建方法，本文提出一种基于引文内容分析的引用情感识别方法，通过采用 tf * idf 结合词性标注的方法构建科技文献引用情感词库，进而提出引用情感判别模型，实现引用情感极性判别。通过对论文中引用情感的准确识别可以为基于引用频次的文献计量提供不同引用行为判定的数据支持。

3 方法框架

为了准确、有效识别出引用情感,在借鉴现有引文动机识别理论与方法的基础上,结合数据挖掘、情感分析和可视化技术,提出一种基于引文内容分析的引用情感识别方法(具体流程见图 1)。该方法以引文全文数据作为研究对象,首先,利用正则表达式技术抽取出论文全文中的引文信息;然后,利用特征词典、情感分析技术识别引用情感;最后,利用可视化分析方法,对引用情感识别结果进行可视化分析,展示引用情感分布情况。

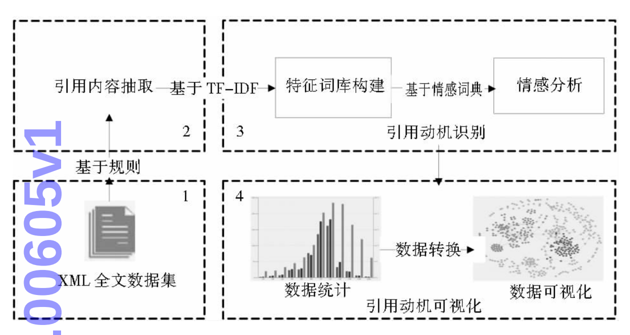


图 1 引用情感识别方法框架

第一步,数据集构建。从相关全文数据库中检索相关领域研究论文,利用 Python 爬虫程序获取 XML 格式的全文数据并保存至本地。

第二步,引文内容抽取。利用 Python 平台编写引文全文信息抽取程序,抽取出施引文献、被引文献的元数据信息和引文内容信息,并写入 CSV 文件用于后续分析。

第三步,引用情感识别。利用 tf * idf 方法筛选出引文全文内容中的特征词,结合情感词典,利用情感分析技术对引用内容进行分析并识别出引用情感。

第四步,引用情感可视化分析。在引用情感识别结果的基础上,构建包含引用情感的复杂网络数据集,揭示出引用情感分布情况。

3.1 引文内容抽取

论文数据格式对引文内容抽取效果影响较大。PDF 格式的全文数据具有不易解析、可读性差等特点,引文内容抽取结果往往准确率较低,而且难以处理大样本的数据。相较于 PDF 等非结构化全文数据,XML 全文数据对论文全文进行了细致标注(对全文数据进行预处理,标注了图表、引用内容和引用位置等信息),便于利用计算机进行大样本数据的引文内容信息抽取、分析。

以 PubMed 数据库中 *Neuroprotective and Anti-Aging Potentials of Essential Oils from Aromatic and Medicinal Plants* 一文 XML 全文文本为例,对其结构进行分析(见图 2),可以发现其 XML 全文文本数据中包括题名、期刊、作者、摘要、图表、引用内容和引用位置等众多标识信息,利用 Python 编写信息抽取程序,可以抽取其中的施引文献著录信息(*<article-title>* … *</article-title>*、*<article-id pub-id-type = "pmid" > 28611658 </article-id>*)、被引文献标签与引文内容(*ref-content*, *Author*, *<xref rid = "" ref-type = "" > pub-date </xref>*)和被引文献标签与著录信息(*<ref-list>* … *</ref-list>*、*<ref id = "B1" >* … *</ref>*)。

如何抽取全文 *ref-content* (*Author*, *<xref rid = "" ref-type = "" >pub-date </xref>*)标签的引文内容是其重点与难点。考虑到作者写论文过程中参考文献序号使用的不规范、不统一情况,构建引文内容抽取规则时需要着重考虑以下两种情况:

(1)只提及一篇参考文献:*The EOs are abundant in flowers, leaves, barks and are usually isolated via hydro-*



图 2 XML 全文结构与引用内容信息抽取

distillation, cold pressing methods (***Edris*, *<xref rid = "B37" ref-type = "bibr" >2007 </xref>***).

句子中只有一处 *ref-content* 标签(一篇参考文献),那么 *ref-content* (*Author*, *<xref rid = "" ref-type = "" > pub-date </xref>*)标签所在整个句子作为引文内容。

(2)提及两篇及两篇以上:*The use of EOs as therapeutic remedy is very ancient and in the bible* (***Dumas and Newhouse*, *<xref rid = "B36" ref-type = "bibr" >2011 </xref>***), *EOs were considered as spiritual, mental and*

physical healing agents（***Guenther***，***<xref rid = "B48" ref-type = "bibr">1950</xref>***）.

Thus, a boost in the cholinergic tone may potentially regress the cognitive function.

句子中有两处及以上 *ref-content* 标签（两篇及以上参考文献），那么以相邻的两个 *<xref rid = "" ref-type = "">pub-date</xref>* 标签作为标记划分引用句，最小单位为一个分句。

3.2 引用情感识别

3.2.1 引用情感类别的划分　半个多世纪以来，E. Garfield、H. Small 和 W. Shadish[21]等众多专家学者针对引用情感类别的划分进行了深入研究。从分析方法角度来看，逐渐由问卷调查分析和人工判读等方法向自然语言处理、情感分析和可视化分析方法转变；从划分类别角度来看，引用情感类别的划分也逐渐随着分析方法的变化而变化，最明显的特征就是划分类别越来越简洁、概括，便于大样本数据的分析。引用类别划分的变化主要有两方面的原因，首先是引用情感识别方法的转变造成难以准确、细致识别出过于复杂的引用情感；其次，由于引文分析数据量的爆炸式增长，过于细致的引用情感类别划分会造成分析效率、准确性的降低。

因此，本文借鉴 S. Teufel、刘盛博等学者的引用情感划分类别[19－20]，将引用情感划分为正面引用、负面引用和中立引用三种类别。

3.2.2 基于特征词和情感词典的引用情感识别　由于期刊论文的规范性、科学性，全文中很少出现感情色彩强烈的语句，而且期刊论文与微博、论坛数据相比，缺少情感词汇，难以通过简单的情感词判断引用情感。因此，仅仅利用情感分析技术进行引用情感识别具有一定局限[22－24]。本文利用 tf＊idf 加词性标注筛选极性特征词和通用情感词典相结合的方法进行引用情感识别。

引文内容主要由内容词和特征词构成。其中，内容词是引文中传递信息的主体，主要以名词形式体现；特征词是引文中表达的情感和状态主要以形容词、动词和连接词形式体现。一个句子中除去名词、介词和限定词等，通过分析特征词可以有效理解句子的情感。因此，首先利用 Stanford POS Tagger[25]进行词性标注，从以下 4 种类型对特征词进行词性标注：形容词（*JJ*）、动词（*VB*）、副词 *RB* 和连接词（*CC*）。例如，对句子 "*The objective of anti-aging medicine is to live as long as possible in good health.*" 进行词性标注，标注结果为：*The_DT objective_NN of_IN anti-aging_ NN medicine_NN is_VBZ to_TO live_VB as_RB long_RB as_IN possible_JJ in_IN good_JJ health_NN. _.*

然后，基于 TF-IDF 算法从中筛选出作为特征词的形容词（*JJ*）、动词（*VB*）、副词 *RB* 和连接词（*CC*）。TF-IDF 算法公式如下：

$$TF\text{-}IDF_{i,j} = TF_{i,j} \times IDF_{i,j}\left(TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}},\ IDF_{i,j} = \log \frac{|D|}{1+|D_{t_i}|}\right) \quad (1)$$

其中，$n_{i,j}$ 指某一词语 t_i 在文件 d_j 中出现的次数，$\sum_k n_{k,j}$ 指文档 d_j 中所有词语数之和；$|D|$ 指文档总数，$|D_{t_i}|$ 指包含词语 t_i 的文档数目（$1+|D_{t_i}|$ 为了防止分母为 0）。

利用 TF-IDF 算法筛选出特征词后，结合 HowNet 构建情感词典并进行情感词汇的赋值（见表 1），如 HowNet 情感词典中没有相应词汇，本文人工进行了标注整理，最终形成引用情感词典。利用公式（2），对引文内容进行判断引用情感，以期提高识别的准确性。

表 1　引用情感与情感词典

引用情感	情感词汇	强度值	值范围
正面引用	accurate丨 better丨 complete 丨 convinced 丨agreement丨 proud……	1	$(0, +\infty)$
负面引用	burden丨contrast丨degrade丨lack丨poor丨worse丨however丨 regret……	－1	$(-\infty, 0)$
中立引用	apply丨 describe丨 discuss丨 publish丨 use……	0	0

情感强度 *E* 的计算方法，如公式（2）所示：

$$E = \frac{1}{n}\sum_{i=1}^{n} E(S_i) \quad (2)$$

其中，*E* 表示引用句的情感强度值（由引用句中各个情感词的情感强度值相加），$E(S_i)$ 表示引用句中某词语 S_i 的情感强度值。

比如，计算某一引用句 "***However***, *this algorithm is very* ***slow*** *and has been* ***outperformed*** *by more recent methods.*" 的情感强度值，基于构建的情感词典可以识别出其中的情感词汇 "***However***"、"***slow***" 和 "***outperformed***"，利用公式（2）进行计算可得该引用句的情感强度值 $E = [(-1) + (-1) + (-1)]/3 = -1$，进而判断该引用为负面引用。

3.2.3 引用情感可视化　由于识别出的引用情感结果无法从整体上表明不同引用情感文献在数据集中的分布情况，所以需要在传统引文网络可视化分析方法

的基础上,利用社会网络可视化软件,如 Gephi[26] 对引用情感识别结果进行可视化分析。引用情感可视化图谱构建的具体步骤为:

(1)数据转换。将引用情感识别结果表格中的施引文献 ID,被引文献 ID 和引用情感三项数据,分别作为 Source、Target 和 Weight 标签构建 Gephi 能够识别的边表格数据形式,其中,正面、负面和中立引用的 Weight 分别标记为 1, -1 和 0,然后保存为. csv 格式文件以备分析。

(2)初始引文网络可视化图谱构建。基于上一步骤生成的边表格数据,导入 Gephi 计算、解析边、节点数据生成初始引文网络图谱,然后自定义图谱的布局(节点、边的大小和颜色等)。

(3)标识引用情感的引文网络可视化图谱构建。在初始引文网络可视化图谱的基础上,基于引用情感数据进一步调整引文网络可视化图谱的布局,进行标识引用情感的引文网络可视化图谱构建,具体设置是将 Weight = 1, -1 和 0 的边赋予红色、绿色和黄色有向线段表示正面、负面和中立引用。

4 实验与分析

4.1 实验环境与数据集构建

(1) 硬件。Windows10 系统、i5 - 2450 CPU、8GRAM、500G Hard Drive。

(2) 软件。Python、KNIME、Stanford POS Tagger、Gephi。

(3) 数据集构建。以 PubMed 生物医学数据库所收录的抗衰老(Anti-aging)领域 XML 格式的论文全文为研究对象。通过检索式:(TITLE:"anti-aging" OR ABSTRACT:"anti-aging" OR KW:"anti-aging"),检索范围:题名,时间跨度:截止至 2016 年 12 月 31 日,对 PubMed 数据库进行检索,检索到 1 135 篇相关论文。利用 PubMed 提供的 OpenURL 接口,编写了 Python 爬虫程序对其 XML 全文数据进行爬取并保存至本地计算机。

基于 3.1 的分析,抽取出施引文献与被引文献的关联关系以及被引文献在全文中所对应的引用内容,共获得 45 527 个引用句,并人工标注了 2 000 个句子的引用极性,见图 3。其中,article-id1 表示施引文献,article-id2 表示被引文献,ref-content 表示引用内容,section 表示引用章节、位置,pub-date 表示被引文献发表时间。

1	article-id1	ref-content	article-id2	section	pub-date
2	1413581	All strains were maintained at 15 on nematode growth medium NGM	4366476	Experimental	1974
3	1413581	By several measures, BB treatment slowed aging in C. elegans , ra	7253717	Results	1981
4	1413581	Regardless of the specific mechanism involved, it is clear from t	3374177	Discussion	1988
5	1413581	Using the short-lived nematode, C. elegans , we have established	8247153	Discussion	1993
6	1413581	To measure levels of 4-HNE, animals were collected, fixed with 4%	8254383	Experimental	1993
7	1413581	Finally, oxidative stress appears to be a major factor limiting l	8415630	Introduction	1993
8	1413581	Lipofuscin levels increase with aging in many organisms, includin	7934213	Results	1994
9	1413581	This mutation results in overproduction of superoxide and increas	7934213	Results	1994
10	1413581	One possible explanation for the beneficial effects of BB polyphe	7638227	Results	1995
11	1413581	We next considered the possibility that BB treatment acted as a r	7638227	Results	1995
12	1413581	We examined lipofuscin levels in the intestines of control and BB	7864834	Results	1995
13	1413581	Interestingly, BB treatment did not improve survival under mild t	9716135	Results	1998
14	1413581	Second, calorie restriction, the only known intervention that suc	9789046	Introduction	1998
15	1413581	In particular, 8-week dietary supplementation with spinach, straw	10479711	Introduction	1999
16	1413581	Studies have shown that specific genetic and environmental factor	10747056	Introduction	2000
17	1413581	One contributor to late-age mortality in C. elegans is the detrin	10747056	Results	2000
18	1413581	To further investigate the effects of blueberries on parameters i	11089983	Introduction	2000
19	1413581	However, only the BB-supplemented group exhibited improved perfor	11099865	Introduction	2000
20	1413581	Blueberries contain a mixture of different polyphenol compounds t	11099865	Results	2000
21	1413581	Treatment with the polyphenol, resveratrol, or related compounds,	11242085	Results	2001
22	1413581	Fractionation of the bulk polyphenols showed that BB's benefits c	11527963	Discussion	2001
23	1413581	Plants synthesize an array of chemical compounds that are not inv	11960739	Introduction	2002
24	1413581	SEK-1/MAPKK is also required for resistance to pathogenic bacteri	12142542	Results	2002
25	1413581	BB treatment also delayed the accumulation of aging-related cellu	12208347	Results	2002
26	1413581	We next examined the effects of BB treatment on a transcriptional	12372248	Results	2002
27	2127228	The absorption, metabolism and tissue distribution of FA has been	13416264	Discussion	1957
28	2127228	Intraperiotneal administration of FA to the rats is excreted as 3	13654297	Discussion	1959
29	2127228	The physiological importance of FA and notably its antioxidant pr	13654297	Discussion	1968
30	2127228	FA is a ubiquitous plant constituent that arises from the metabol	11121513	Discussion	1983

图 3　引用内容信息表格(部分)

4.2 引用情感识别

4.2.1 基于 TF-IDF 的特征词筛选　引用内容信息抽取完成后,利用 Stanford POS Tagger 工具标注出引文内容的词性,然后基于 TF-IDF 筛选出其中权重较高的特征词(形容词 *JJ*、动词 *VB*、副词 *RB* 和连接词 *CC*)(部分结果见图 4)。对 45 629 个引用句进行词性标注,共标注出 1 197 191 词语的词性,利用数据挖掘软件 KNIME 分别计算其 TF、IDF 和 TF-IDF 值,然后分别统计形容词(*JJ*)、动词(*VB*)、副词 *RB* 和连接词(*CC*)得到基于 TF-IDF 值的特征词列表(见表 2)。

4.2.2 基于情感词典的引用情感识别　结合基础词典 Hownet 中的情感词汇和筛选出的特征词,构建了抗衰老领域情感词典。具体操作步骤为:首先,将筛选出的抗衰老领域特征词进行人工标注,分别标注 TF-IDF 值排名前 300 的形容词(*JJ*)、动词(*VB*)、副词 *RB* 和连接词(*CC*),并将其划分为正面、负面和中立三类;然后,与 Hownet 中的情感词汇相结合,对两个词表中的词汇进行去重、合并;最后得到本实验中需要的抗衰老领域的情感词典。其中,正面情感词汇共有 4 570 个,负面情感词汇共有 4 363 个,中立情感词汇共有 235 个,见图 5。

根据本文提出的引用情感计算模型,在数据挖掘平台 KNIME 上,利用人工标注的引用极性数据集合进行了情感极性判别实验。实验分两组进行,实验 1 组利用基础词典 Hownet 中的情感词汇作为特征词,实验 2 组利用本文构建的情感词典作为特征词。利用本文 3.2.2 提出的情感判别方法,将抽取出的 Pubmed 中抗衰老领域 45 527 个引用句通过 KNIME 平台分别根据 2 组不同词典进行了实验。实验流程见图 6。

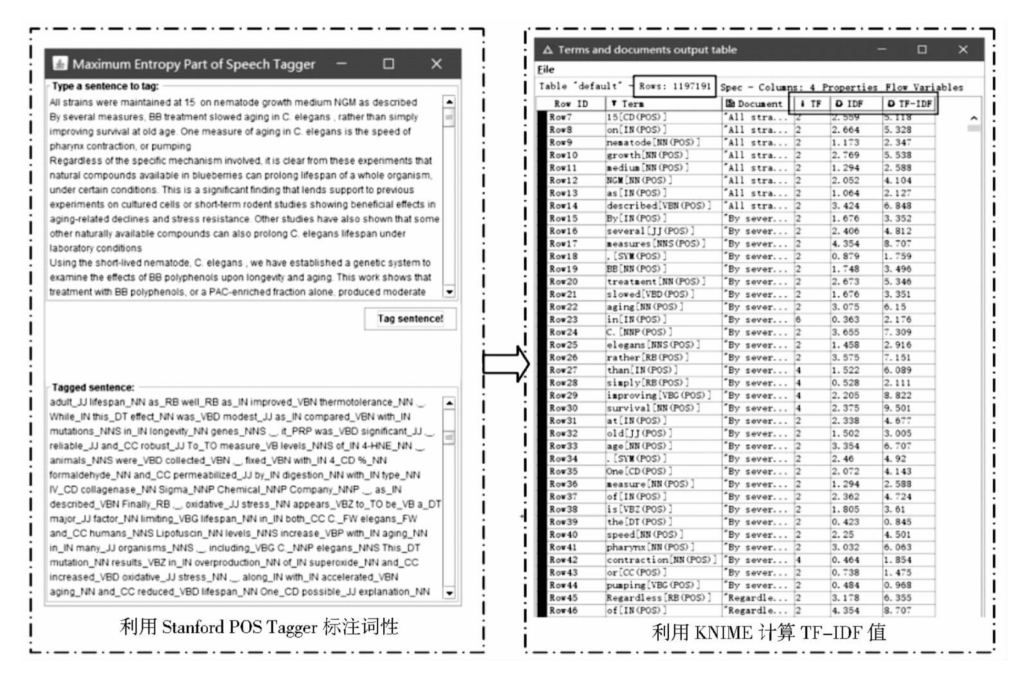


图 4　词性标注与 TF-IDF 值计算

表 2　基于 TF-IDF 值的特征词列表(部分)

No	形容词	TF-IDF	动词	TF-IDF	副词	TF-IDF	连接词	TF-IDF
1	Stretchy	83.55	Degrade	26.20	Alone	19.45	And	16.61
2	Heavy	46.51	Wound	23.73	Ectopically	18.62	Or	11.19
3	Radical	44.51	Reverse	22.74	Firmly	15.01	Both	10.51
4	Magnetic	45.84	Flash	18.62	Adequately	15.01	However	9.23
5	Phenoxy	37.24	Touch	17.41	Additionally	14.24	Yet	8.95
6	Immune	36.27	Count	17.41	Only	14.03	If	8.63
7	Forward	33.42	Discuss	15.51	Accurately	13.70	Still	8.31
8	Embryonic	31.97	Exclude	14.80	Tightly	12.09	Since	8.31
9	Intronic	29.24	Report	14.30	Mostly	10.52	Because	8.31
10	Pranic	27.93	Follow	13.41	Partially	10.52	While	7.40
…	…	…	…	…	…	…	…	…

Output table - 2:294 - XLS Reader

File

Properties　Flow Variables

Table "抗衰老领域情感词典.xlsx [Sheet1]" - Rows: 4570　Spec - Columns: 3

Row ID	S 正面	S 负面	S 中性
Row12	ably and efficiently	above standard	by all means
Row13	abound in gifts of n...	above the commo...	completely
Row14	above criticism	abrupt	deep-rooted
Row15	above-board	abruptly	deep-seated
Row16	aboveboard	abruptness	deeply
Row17	absolutely fair	absent-minded	definitely
Row18	absolutely fearless	absent-mindedly	disastrously
Row19	absolutely necessary	absentminded	downright
Row20	absolutely true	absolutely irre...	entirely
Row21	absorbed	absolutely lawl...	exceedingly
Row22	abstemious	absolutely vici...	excessively
Row23	abstemiousness	absolutely wrong	extreme
Row24	abstruse	absorption	extremely
Row25	abundant	absurd	fully
Row26	abundantly	absurdist	greatest
Row27	abundantly clear	absurdity	greatly
Row28	accessibility	absurdly	heinous
Row29	accommodating	acclivous	hundred-per...
Row30	accordant	accursed	immensely
Row31	according to reason	acedia	immoderate
Row32	according to rules	acerb	in a penetr...
Row33	according to the facts	acerbic	in every po...
Row34	accurate	acerbity	in the extreme
Row35	accurately	acid	incomparably
Row36	ace	acidulous	ingrained
Row37	act as the occasion ...	acquisitive	matchlessly

图 5　抗衰老领域情感词典

图书情报工作
第 62 卷 第 15 期 2018 年 8 月

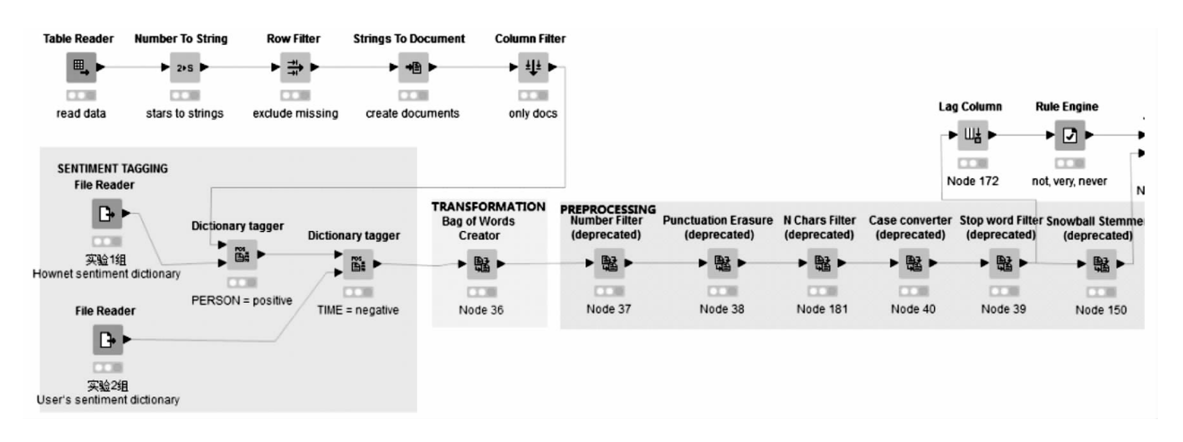


图 6 两组实验过程

在结果评估方面，采用精度（Precision）、召回率（Recall）和 F1 值 3 个指标。实验结果如表 3 所示：

表 3 实验结果分析

	精度（P）		召回率（R）		F1	
	实验 1 组	实验 2 组	实验 1 组	实验 2 组	实验 1 组	实验 2 组
正面引用	77.31%	78.62%	79.32%	80.13%	78.30%	79.37%
负面引用	81.20%	83.00%	83.20%	84.00%	82.18%	83.34%
中立引用	76.16%	76.31%	75.36%	76.19%	75.75%	76.25%

从表 3 可以看出实验 2 组在精度、召回率以及 F1 值均有不同程度的性能提升，特别是在负面引用判别上，性能提升更加明显。说明本文构建的引用情感词典可以有效提高识别效果。实验 2 组方案实验结果中，正面引用占总引用次数的 21%，中立引用占总引用次数的 78%，负面引用仅占总引用次数的 1%，如表 4 所示：

表 4 引用情感百分比统计表

引用情感	引用次数	百分比
正面引用	2 452	20.74%
负面引用	169	1.43%
中立引用	9 196	77.82%

图 7 显示了 KINME 平台中的具体实验结果，其中正面引用（Positive，POS）、负面引用（Negative，NEG）和中立引用（Neutral，NEU）分别设置为红色、绿色和黄色。

Table with Colors - 4:310 - Color Manager (Color by sentiment)
File
Table "default" – Rows: 45527 Spec – Columns: 6 Properties Flow Variables

Row ID	article-id1	ref-content	article-id2	P (Document class=NEG)	P (Document class=POS)	P (Document class)
Row0	1413581	All strains...	4366476	0.986	0.014	NEG
Row1	1413581	By several ...	7253717	1	0	NEG
Row2	1413581	Regardless ...	3374177	0.987	0.013	NEG
Row3	1413581	Using the s...	8247153	0	1	POS
Row4	1413581	To measure ...	8254383	0.003	0.997	POS
Row5	1413581	Finally, ox...	8415630	0	1	POS
Row6	1413581	Lipofuscin ...	7934213	0.986	0.014	NEG
Row7	1413581	This mutati...	7934213	0.003	0.997	POS
Row8	1413581	One possibl...	7638227	1	0	NEG
Row9	1413581	We next con...	7638227	0.987	0.013	NEG
Row10	1413581	We examined...	7864834	1	0	NEG
Row11	1413581	Interesting...	9716135	0.986	0.014	NEG
Row12	1413581	Second, cal...	9789046	0.991	0.009	NEG
Row13	1413581	In particul...	10479711	0.9	0.1	NEG
Row14	1413581	Studies hav...	10747056	0	1	POS
Row15	1413581	One contrib...	10747056	0.991	0.009	NEG
Row16	1413581	To further ...	11089983	0.987	0.013	NEG
Row17	1413581	However, on...	11099865	0.003	0.997	POS
Row18	1413581	Blueberries...	11099865	0.056	0.944	POS
Row19	1413581	Treatment w...	11242085	0.003	0.997	POS
Row20	1413581	Fractionati...	11527963	0.003	0.997	POS
Row21	1413581	Plants synt...	11960739	0	1	POS
Row22	1413581	SEK-1/MAPKK...	12142542	0.003	0.997	POS
Row23	1413581	BB treatmen...	12208347	0.986	0.014	NEG
Row24	1413581	We next exa...	12372248	0.003	0.997	POS
Row25	2127228	The absorpt...	13416264	0.003	0.997	POS
Row26	2127228	Intraperiot...	13654297	0.003	0.997	POS
Row27	2127228	The physiol...	13654297	0.987	0.013	NEG
Row28	2127228	FA is a ubi...	11121513	0.003	0.997	POS
Row29	2127228	have report...	3555751	0.003	0.997	POS

图 7 抗衰老领域引用情感识别结果

4.3 引用情感可视化

本文提出的引用情感可视化分析是在传统引文网络可视化图谱的基础上，添加引用情感标记，构建出标识引用情感的引文网络图谱，从而可以有效发现不同引用情感在整体数据集上的分布情况。

将被引次数为零的节点剔除（剩余节点 1127，边 1191），处理后得到抗衰老领域——初始引文网络，见图 8。图 8 中，蓝色节点表示论文，节点大小正比于被

引次数，节点标签为论文 PMID（PubMed 生物医学数据库为每篇论文赋予唯一 ID），为了优化图谱显示效果（如果显示所有节点标签 ID 会遮挡其它关键信息），图中仅展示了被引次数排名前 20 论文节点的 ID；黄色有向连线表示引用方向，A→B 表示 B 引用 A。

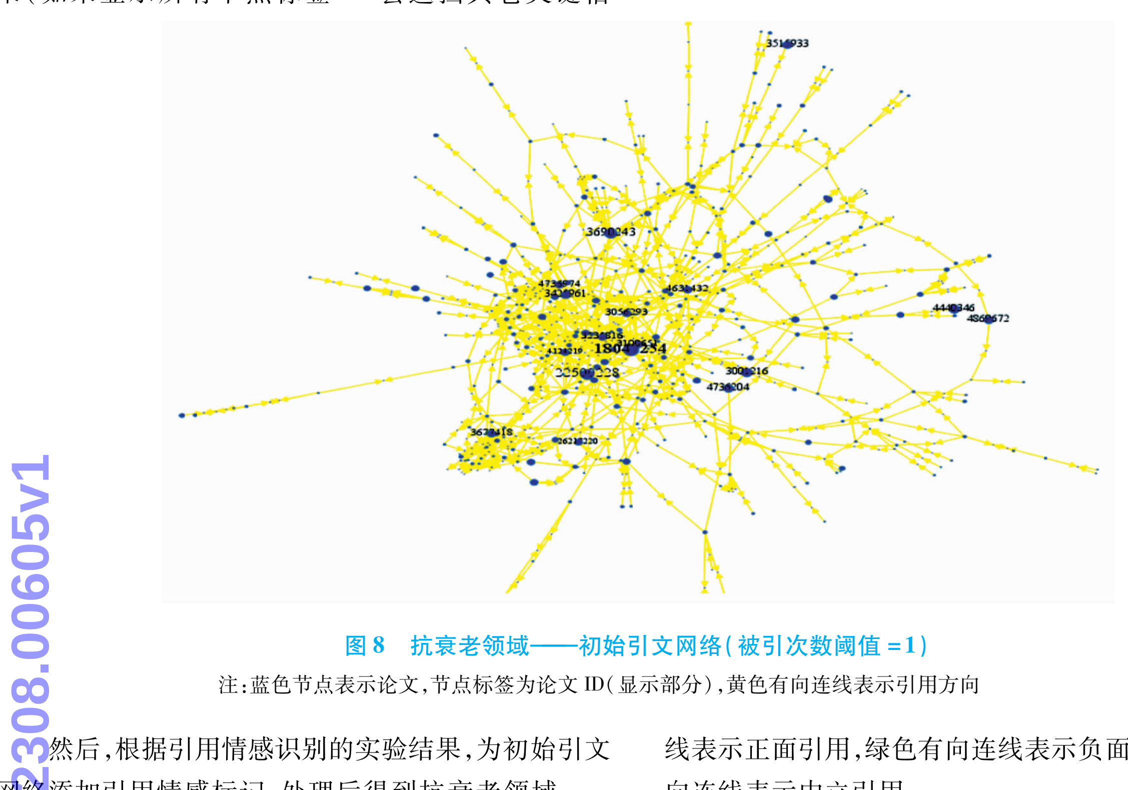


图 8　抗衰老领域——初始引文网络（被引次数阈值 =1）

注：蓝色节点表示论文，节点标签为论文 ID（显示部分），黄色有向连线表示引用方向

然后，根据引用情感识别的实验结果，为初始引文网络添加引用情感标记，处理后得到抗衰老领域——标识引用情感的引文网络，见图 9。其中，红色有向连线表示正面引用，绿色有向连线表示负面引用，黄色有向连线表示中立引用。

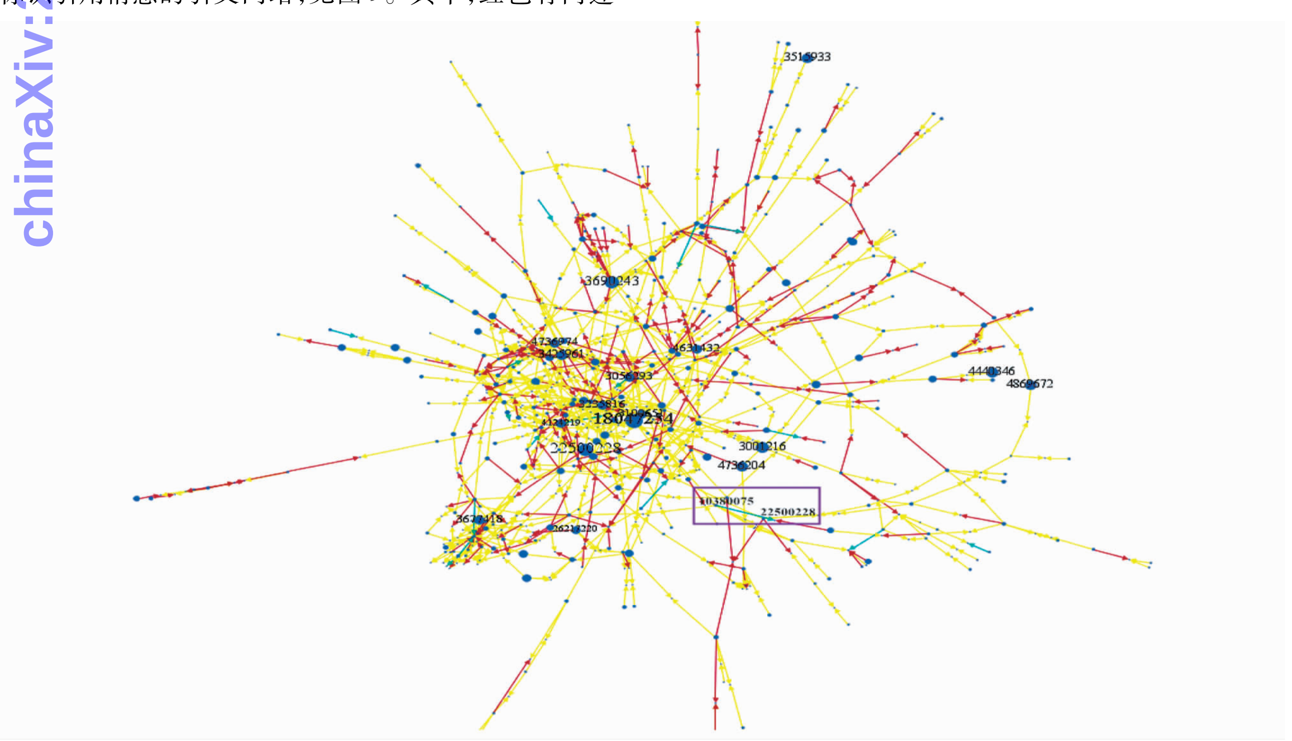


图 9　抗衰老领域——标识引用情感的引文网络（被引次数阈值 =1）

注：蓝色节点表示论文，节点标签为论文 ID（显示部分），有向连线表示引用方向与引用情感，其中，红色表示正面引用，绿色表示负面引用，黄色表示中立引用

与传统引文网络相比较,图 8 可以有效识别出不同引用情感在整体数据集合上的分布情况。可以进一步分析正向引用链,负面引用链,为基于引文内容分析的文献计量学提供支持。例如:节点 10380075 存在一条负面引用(10380075→22500228,图 8 中紫色方框标出),为了验证该引用情感识别结果的准确性以及可视化图谱的有效性,在 PubMed 数据库中,分别基于 PMID = 22500228 和 PMID = 10380075 检索得到文献 *Insulin in central nervous system: more than just a peripheral hormone* 和 *Phosphatidylinositol 3-kinase-mediated regulation of neuronal apoptosis and necrosis by insulin and IGF-I*,然后定位至引用内容:

In contrast, Ryu et al. [160] failed to show protection by IGF-1 against excitotoxic or oxidative stress-induced necrosis, despite a decrement in neuronal apoptosis.

通过人工判读引文内容"相比之下,尽管在神经细胞凋亡减少,Ryu 等人未能证明 IGF-1 能够防护、避免细胞发生兴奋性毒性或氧化应激引起的细胞坏死",可以知道该引用情感是负面引用,即通过引用 Ryu 等人的研究来说明、反衬研究者发现胰岛素、IGF-1 能够弱化视网膜和脑神经元凋亡的条件(如氧化应激反应等)这一研究的价值和创新。因此,该引用情感识别结果是准确的,同时也证明了引用情感可视化分析的可行性和有效性。

5 结语

基于引文内容分析的引用情感识别相关研究一直是引文分析研究领域的研究热点,特别是近年来随着全文数据库和数据挖掘技术的发展,以及基于引用次数的科研评价体系受到质疑,如何利用现代信息技术高效、准确识别引文内容情感并进行可视化分析,有待研究者进行深入研究。

实验结果表明,本文提出的基于文本内容分析的引用情感识别方法与目前研究中的引用情感分析方法相比,一方面构建了基于特征词和基础词典的抗衰老领域专门情感词典,提高了引用情感识别结果的准确性;此外,进一步提出了相应的引用情感可视化图谱构建方法,有效识别出不同引用情感在整体数据集合上的分布情况,在一定程度上增加了该研究的应用价值。

本文将引用情感分为了正面、负面和中立引用三种,而实际科研活动中引用行为、动机更加复杂。因此,在下一步的研究中,将探索如何识别不同引用情感的重要程度,以及引用方法、引用工具、引用模型等更加细粒度的引用情感的识别。

参考文献:

[1] 赵蓉英,王建品. 引用内容分析与引文著录分析的比较研究[J]. 图书情报工作,2017,61(10):110-115.
[2] 国家科学技术奖励工作办公室. 国家自然科学奖奖励介绍[EB/OL]. [2017-07-27]. http://www.nosta.gov.cn/web/detail.aspx? menuID=158&contentID=1115.
[3] 胡志刚,陈超美,刘泽渊,等. 基于 XML 全文数据引文分析系统的设计与实现[J]. 现代图书情报技术,2012(11): 72 - 77.
[4] MORAVCSIK M J, MURUGESAN P. Some results on the function and quality of citations[J]. Social studies of science, 1975, 5(1): 86-92.
[5] SMALL H G. Cited documents as concept symbols[J]. Social studies of science, 1978, 8(3):327-340.
[6] SMALL H G, GREENLEE E. Citation context analysis of a co-citation cluster: recombinant-DNA[J]. Scientometrics, 1980, 2(4): 277-301.
[7] DING Y. Content-based citation analysis: the next generation in citation analysis[EB/OL]. [2012-09-26]. http://www.lis.illinois.edu/events/2012/09/26/content-based-citation-analysis-next-generation-citation-analysis.
[8] 祝青松,冷伏海. 基于引文内容分析的高被引论文主题识别研究[J]. 中国图书馆学报,2014,40(1):30-49.
[9] 陆伟,孟睿,刘兴帮. 面向引用关系的引文内容标注框架研究[J]. 中国图书馆学报,2014,40(6):93-104.
[10] 赵蓉英,曾宪琴,陈必坤. 全文本引文分析-引文分析的新发展[J]. 图书情报工作,2014,58(9):129-135.
[11] 赵蓉英, 郭凤娇, 曾宪琴. 基于位置的共被引分析实证研究. 情报学报,2016,35(5):492-500.
[12] GARFIELD E. Can citation indexing be automated? [J]. Essays of an information scientist, 1962, 1: 84-90.
[13] CANO V. Citation behavior: classification, utility, and location [J]. Journal of the American Society for Information Science, 1989,40(4): 284-290.
[14] LIU M X. Progress in documentation the complexities of citation practice: a review of citation studies[J]. Journal of documentation, 1993, 49(4):370-408.
[15] CASE D O, HIGGINS G M. How can we investigate citation behavior? A study of reasons for citing literature in communication [J]. Journal of the American Society for Information Science, 2000, 51(7):635-645.
[16] KESSLER M M. Bibliographic coupling between scientific papers [J]. American documentation wiley online library, 1963, 14(1): 10-25.
[17] SMALL H G. Co-citation in the scientific literature: a new measure of the relationship between two documents[J] Journal of the American Society for Information Science, 1973, 24 (4): 265-269.
[18] CHUBIN D E, MOITRA S D. Content analysis of references: adjunct or alternative to citation counting? [J]. Social studies of sci-

ence, 1975, 5(4): 423－441.

[19] TEUFEL S, SIDDHARTHAN A, DAN T. Automatic classification of citation function[C]//Conference on empirical methods in natural language processing, 出版地:出版者 2006, 14 (1):103－110.

[20] 刘盛博,丁堃,张春博. 基于引用内容性质的引文评价研究[J]. 情报理论与实践,2015,38(3):77－81.

[21] SHADISH W R, TOLLIVER D, GRAY M, ET AL. Author judgments about works they cite: three studies from psychology journals [J]. Social studies of science, 1995, 25(3): 477－498.

[22] Verlic M, Stiglic G, Kocbek S, et al. Sentiment in science a case study of cbms contributions in years 2003 to 2007[C]//Computer-based medical wystems, 2008 CBMS '08 21st IEEE international symposium on. albuquerque, Jyväskylä, Finland: IEEE, 2008: 138－143.

[23] SMALL H. Interpreting maps of science using citation context sentiments: a preliminary investigation[J]. Scientometrics, 2011, 87 (2): 373－388.

[24] BONACICH P B. Factoring and weighting approaches to status scores and clique identification[J]. Journal of mathematical sociology, 1972, 2 (1):113－120.

[25] TOUTANOVA K, MANNING C D. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger[C]// Joint sigdat conference on empirical methods in natural language processing, Hong Kong: ACM, 2000, 25 (6):63－70.

[26] JACOMY M, BASTIAN M, HEYMANN S. Gephi: an open source software for exploring and manipulating networks[C]// International conference on weblogs & social media. San Jose: The AAAI Press, 2009:361－362.

作者贡献说明:

廖君华:进行研究命题和思路设计;
刘自强:负责数据分析和论文撰写;
白如江:负责框架制定和数据收集;
陈军营:负责数据收集和预处理。

Citation Sentiment Recognition Method Based on Citation Content Analysis

Liao Junhua[1] Liu Ziqiang[2,3] Bai Rujiang[1] Chen Junying[1]

[1] Institute of Scientific Technical Information, Shandong University of Technology, Zibo 255049

[2] Chengdu Library and Information Center, Chinese Academy of Sciences, Chengdu 610041

[3] University of Chinese Academy of Sciences, Beijing 100190

Abstract: [**Purpose/significance**] The paper proposes an identification method based on the analysis of citations content. And a visual display is presented to overcome the problem of different citation emotions based on simple reference frequency measurement. [**Method/process**] First, it uses regular expressions to extract the content information of the text in full text. Then, it uses the TF-IDF algorithm to select the quoted emotion feature words, combines the emotional dictionary, and uses emotional analysis technology to quote emotion recognition. Finally, the use of visual tools shows the overall distribution of the reference emotion. [**Result/conclusion**] The method can effectively identify emotional information in the domain of anti-aging. The experimental results show that the positive citation accounts for 21% of the total citation frequency, neutral citation accounts for 78% of the total citation frequency, and negative citation accounts for only 1% of the total citation frequency. Compared with the traditional citation network, the visualization map based on citation emotion can effectively identify the distribution of different citation emotions on the overall data set.

Keywords: citation content analysis citation motivation emotion analysis visualization